# PageX: An Integrated Document Processing and Management Software for Digital Libraries

Hanchuan Peng, Zheru Chi, Wanchi Siu, and David Dagan Feng

Department of Electronic & Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Email: phc@eie.polyu.edu.hk

This paper is submitted for MMWS2000.

Category of this paper should be "Multimedia Applications".

Author in correspondence should be:

Dr. Hanchuan Peng
EIE,
The Hong Kong Polytechnic University,
Hong Kong.
Email: phc@eie.polyu.edu.hk

# PageX: An Integrated Document Processing and Management Software for Digital Libraries

Hanchuan Peng, Zheru Chi, Wanchi Siu, and David Dagan Feng

Department of Electronic & Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Email: phc@eie.polyu.edu.hk

***Abstract* – For digital libraries it is very important to design and to implement powerful engines to convert information on paper to electronic format. In this paper a PageX software is proposed as the integration of such engines, which include a set of intelligent document processing functions, a set of compact document management strategies, and a set of advanced accessories. With this software, a paper document will first be input as an optical image, which may be a mixture of graphics and text and may be skewed. The image will then be analyzed and decomposed into a series of component blocks, and encoded and stored in a structured and compact format. Well-developed accessory functions, including block editing and page annotation, page reconstruction and virtual editing, page matching and registration, document retrieval, *etc.*, are provided to support advanced applications. With these carefully designed functions and strategies, PageX minimizes manual operations to a minimal degree.**

***Index Terms* – Digital library, document processing, document database, page segmentation**

## I. Overview of PageX System

Digital libraries are some of the most typical applications of multimedia data storage and retrieval [1-3]. Page document oriented software [4] is an important aspect of digital library research. It has been emphasized that high-accuracy Optical character recognition (OCR) is greatly required in document auto-reading systems. However, OCR does not perform a unique role in document processing. Actually, due to a series of unavoidable factors (such as a mixture of text and graphics in documents, shadow and distortion in document image acquiring, *etc.*), very possible the scanned documents can only be treated as raw images, where no significant text information can be extracted from such page images. Clearly, such a case is unbearable for electronic document transmission and storage. Is there any solution to deal with a general document on paper and to automatically convert its scanned image to electronic format and to offer advanced properties of editing and retrieval? Is there any method to decrease the manual work in document processing to a trivial level? This paper introduces PageX as one of our efforts for tackling such problems.

Fig.1 shows the main paradigm of PageX, which consists of three major modules and a TWAIN compatible image acquiring module. The first major module (engine set I) of PageX is intelligent document processing, which automatically analyzes page information and decomposes a page image into a Component block list (CBL). This module has little manual operation. The second module (engine set II) of PageX is compact document management, which encodes all the document components (blocks and other document properties) and offers a flexible data structure for transmission and storage. With the second module, CBL is further organized as compact Electronic document (e-Doc). The third module (engine set III) of PageX is a set of advanced accessory functions for document-oriented applications, such as the editing and output (saving as other formats, printing, dynamically linking and outputting to database, *etc.*). Retrieval engines are also included in this part. These three parts are explained in subsequent sections.

It is interesting to compare PageX with some other document processing research in digital libraries. For example, the document processing system developed by the group of Document image decoding (DID)[*] at Xerox aims at developing Advanced structured documents (ASD) that provide contented-based access to information extracted from the scanned page, but are not necessarily intended to be reconstruction of corresponding paper documents [3]. This is in contrast with PageX, which attempts to provide both structured documents and good reconstruction of the original paper documents. Therefore PageX can avoid

[*] http://elib.cs.berkeley.edu/kopec/

improper or wrong processing, and can be further utilized in areas other than digital libraries, such as document communication in teleconferencing.

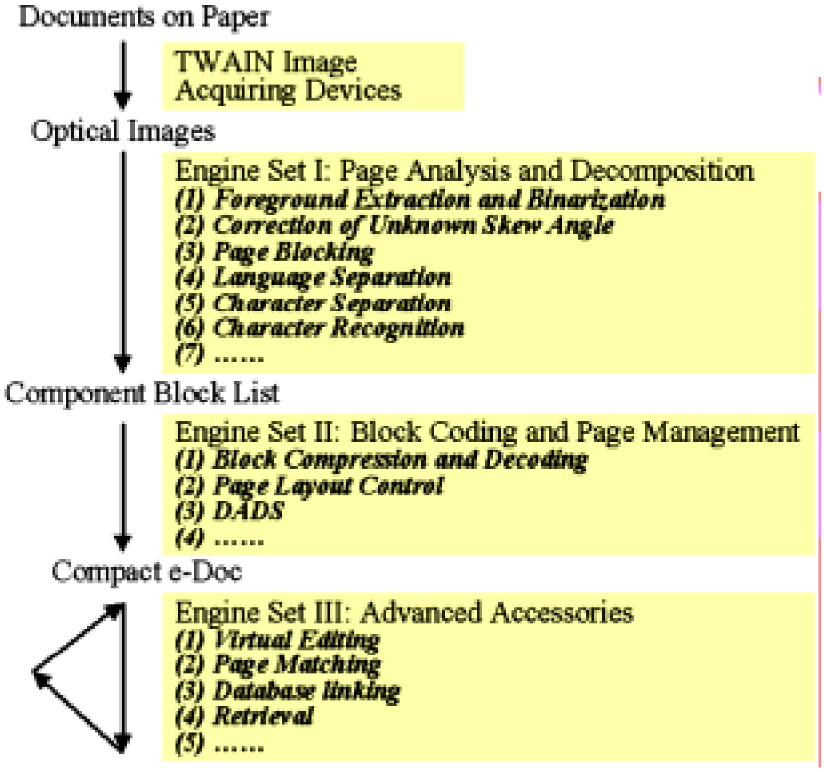

Documents on Paper
TWAIN Image Acquiring Devices
Optical Images
Engine Set I: Page Analysis and Decomposition
(1) Foreground Extraction and Binarization
(2) Correction of Unknown Skew Angle
(3) Page Blocking
(4) Language Separation
(5) Character Separation
(6) Character Recognition
(7) ……
Component Block List
Engine Set II: Block Coding and Page Management
(1) Block Compression and Decoding
(2) Page Layout Control
(3) DADS
(4) ……
Compact e-Doc
Engine Set III: Advanced Accessories
(1) Virtual Editing
(2) Page Matching
(3) Database linking
(4) Retrieval
(5) ……

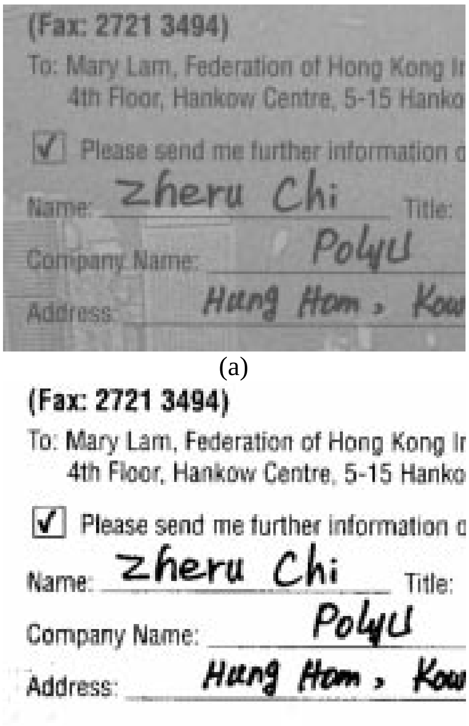
Fig.1 The scheme of PageX

## II. INTELLIGENT DOCUMENT PROCESSING

A set of intelligent functions for document analysis is included in PageX to analyze and extract the page information (Fig.1). These functions include automatic correction for skew images with unknown distortion angles, automatic extraction of foreground from complex background, automatic page blocking, automatic language separation and character segmentation, character recognition, *etc*. With these functions, an acquired (either scanned or captured) page image, which contains different languages of texts, various types of graphics, and other information (Fig.2(a) and Fig.3(a)), will be decomposed into a series of component blocks, which have different types. These blocks make up the CBL of the original document page. The main idea in designing these functions is to provide intelligent/automation techniques to decrease the manual operation as much as possible. Several most important functions are briefly introduced in the following subsections.

### *A. Foreground Extraction and Binarization*

It is unusual for many document reading and processing software packages to include a powerful and intelligent binarization module. Actually, due to the uneven property of input images, a global threshold, either automatically or manually selected, generally leads to poor binarization, in which situation the resulted page image often contains much wrong information. What's worse, it is usually hard to get rid of such wrong information (*e.g.* noisy blocks and irregularly shaped spots). Another problem is the original scanned page images have a strong likeness to comprise foreground texts and background pictures. A typical example is shown in Fig.2(a), where the foreground texts should be extracted from the background picture of a building. A self-adaptive binarization engine is implemented to extract the foreground and to set the background to be white. This engine will firstly analyze all sub-regions of the whole image and then produce a mesh of binarization threshold. This mesh can lead to better binarization outcomes than a global binarization threshold. After binarization, an engine for noise elimination is used to erase most point noise and texture background. The processed image, as shown in Fig.2(b), is seen in good quality.



(Fax: 2721 3494)
To: Mary Lam, Federation of Hong Kong Ir
4th Floor, Hankow Centre, 5-15 Hanko
☑ Please send me further information c
Name: zheru Chi Title:
Company Name: PolyU
Address: Hung Hom, Kow

(a)

(Fax: 2721 3494)
To: Mary Lam, Federation of Hong Kong Ir
4th Floor, Hankow Centre, 5-15 Hanko
☑ Please send me further information c
Name: zheru Chi Title:
Company Name: PolyU
Address: Hung Hom, Kow

(b)

Fig.2 Intelligent binarization and foreground extraction. (a) Input image (grayscale) (b) Binarized image

### *B. Skew Image Correction*

Usually, document pages on paper are scanned with unknown distortion angles, which results in skew images. One example is shown in Fig.3(a). An image rotation engine is implemented to attack this problem. In this engine, a correlation-based method is employed as the first step to automatic detect the unknown skew angle, which can find out a distortion angle up to 45°, even in a non-text page which mainly contain lines. Then, an interpolation method is used to rotate the image with the skew angle smoothly. Fig.3(b) is the automatically rotated image from Fig.3(a). With this image rotation engine, only a very small distortion angle may exist after correction. This small distortion is further removed with a shift-compensation method.

### *C. Page Blocking*

The engine of page blocking is implemented to decompose a page image into rectangular component blocks, each of which is further decided to have one attribute of text, or graphics, line, *etc.* Proper organization of these blocks implies a good solution to reconstruct the original page in the electronic format. Various methods have been proposed to segment the page image. PageX utilizes a block expanding method, in which a block does not grow as large as possible only if there is no more connecting neighbor. The page blocking result of Fig.3(b) is shown in Fig.3(c), where it is seen that all texts and graphics (pictures and logos) are correctly segmented. This engine of page blocking also automatically labels and removes all horizontal and vertical lines.

### *D. Language Separation*

An engine of language separation is implemented to distinguish Chinese texts from English texts. A hybrid algorithm made up of a concavity-based method and a momentum-based method. Chinese blocks often have much more concavities. Similarly, several types of statistical momentum of Chinese scripts are very different from the English scripts. This engine also decides whether a graphic block is a picture or a logo. As an example, the language separation result is shown in Fig.3(d) and (e), which are English parts and Chinese parts, respectively.

### *E. Character Segmentation and Recognition*

A simple but useful engine for character segmentation is also implemented to cut a text block into pieces of character sub-blocks. Typical results are shown in Fig.3(f). In the case of character connection, this engine has not been implemented for a purpose of absolutely error-free in character segmentation. On the contrary, the information produced by this engine, that is, the list of positions of characters, will be very useful for character recognition and subsequent editing purposes. After character segmentation, character recognition is performed to translate most of the text blocks into ASCII characters. The unrecognizable blocks, such as noisy text blocks and graphics blocks, are sent to next step for further processing.

## III. COMPACT DOCUMENT MANAGEMENT

Through the technique of compact document management, PageX constructs an e-Doc consisting of all page component blocks. A set of coding and decoding engines are implemented to compress and decompress blocks. A page layout management engine is designed to reconstruct and handle the document page. A Dynamic adjustable data structure (DADS) [5] is employed to organize all the blocks in a structured manner. Detailed report of the used techniques is given elsewhere. Some of these techniques are concisely introduced as the following.

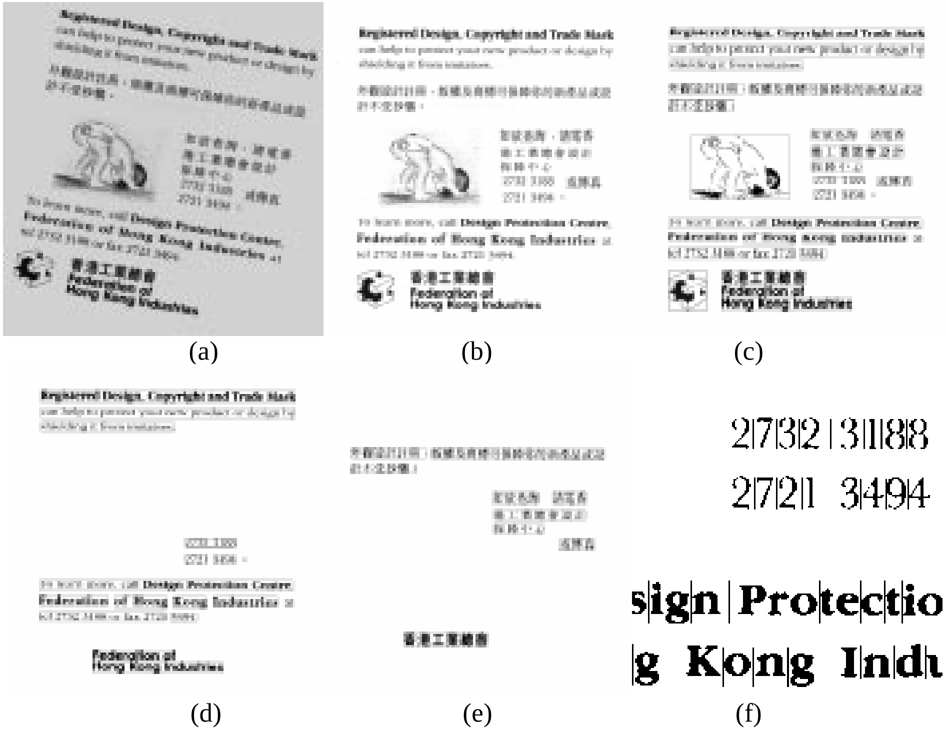

(a)
(b)
(c)
2732 3188
2721 3494
sign Protectio
g Kong Indu
(d)
(e)
(f)

Fig.3 Intelligent document analyzing functions. (a) The input skew image (grayscale) (b) Upright image after correction (c) Results of page auto-blocking (d) English part resulted from the language separation engine (e) Chinese part resulted from the language separation engine (f) Partial results of the character segmentation

*A. Block Compression/Decompression*

After intelligent processing, the following set of information of a block is obtained: {type, position, referencing text, block image data, character sub-block position list}. "Type" of a block is the content of the block: it is a block of text or graphics, and in more details, an English block or Chinese block, and so on. "Position" is the border position of the block in the page image. "Referencing text" is the recognized text of the block, if applicable. "Block image data" is the block's image in the page area bordered by "position". "Character sub-block position list" (CPL) is the position list given by the engine of character segmentation. In the engine of block compression/decompression, "block image data" is compressed with a lossless compression engine. At this time, this engine can compress block to Portable network graphics (PNG) format. Other compression algorithms will be added later.

*B. Page Layout Management*

Attributes of the page layout include all positions of blocks and lines (which are removed by the engine of page blocking in the last section), and image properties of the scanned page (such as the resolution in scanning and the original scanned page size). With such information, the page layout can be synthesized after all blocks' information is decoded. The engine of layout management is implemented to reconstruct the page. This engine also serves as page layout control, which can handle different requests of page layout and realize conversion of various page layouts.

*C. DADS and Flexible e-Doc*

PageX system utilizes DADS [5], which has been designed for easy retrieval of image and video, in data management. DADS provides a flexible data-structure where the polymorphous block can be dynamically linked to diversified action engines. Through DADS, it is easy to realize block/page editing, block/page retrieval, and other operations, in an adaptable manner. For example, for graphics blocks, PageX can quickly locate logos, or other pictures in the scanned page image. Text image data can also be retrieved. Still there is another type of document retrieval being supported: the referencing text retrieval.

## IV. Advanced Accessories

Based on the information produced from the previous two engine sets, several very useful functions have been included in PageX to support advanced applications. These advanced accessories include block editing and page annotation, page reconstruction and virtual editing, page matching and registration, document retrieval, database dynamically linking, and so on. With these additional functions, users can manipulate compact e-Doc easily and effectively. Interfaces of e-Doc for building document databases are also provided.

## V. Conclusion and Prospects

PageX has been devised and implemented as a document-oriented processing and management system. Many intelligent functions are provided to minimize the manual operations and to make the unavoidable manual work as natural and unobservable as possible. By the mean time, the obtained e-Doc has a very compact manner in data storage and transmission. That is, the e-Doc is an easy and accurate reconstruction of the original page document, while only a minimal storage space is required. When the document is needed to change a format, the referencing text and original block formats are used for document reformatting. The transparent and natural way to manipulate the e-Doc is supported by PageX. In total, we see this PageX system in fact offers a new strategy to treat the data reading problem in digital libraries.

PageX is an evolving system. More functions, *e.g.*, engines, are being added to PageX. This software package is expected to be an extensible solution to document reading problem in digital libraries. It is also expected to be useful in e-Doc transmission and sharing.

## Acknowledgement

The authors thank Dr. Fuhui Long for suggestive discussion.

## References

[1] I.H., Witten, R.J., McNab, S., Jones, M., Apperley, D., Bainbridge, and S.J., Cunningham "Managing complexity in a distributed digital library," Computer, Vol.32(2), pp.74-79, 1999.

[2] S.T.C., Wong, and D.A., Tjandra, "A digital library for biomedical imaging on the Internet," IEEE Communications Magazine, Vol.37(1), pp.84-91, 1999.

[3] G.E., Kopec, "Document image decoding in the Berkeley digital library," Proc of 1996 Int Conf on Image Processing, Vol.1, pp.769-772, 1996.

[4] Q., Gan, H.-C., Peng, *et al*, "Research on auto-reading systems of handwritten data forms," Certificate of Scientific/Technical Accomplishment (awarded by the Committee of Science & Technology, Jiangsu Province, China), No. SKJ1996-331, Dec., 1996.

[5] F.-H., Long, D., Feng, *et al*, "Dynamically Adjustable Data Structure for Video Databases," to appear in ISSPIS'99.